期末報告——評論&分數的關係

一、研究動機

在生活當中，人們常常會嘗試一些新事物，比如餐廳、旅遊等，而在這個時候，我們往往會參考他人的意見，最常參考的就是 google 地圖的評論系統或者部落格的評分，而這些呈現往往會有「分數」作為一個更直接的參考，但是，在生活中其實有很多的「評論」是沒有分數作為參考的，分數的確是一個比較直觀的方法，因此想從由文字構成、較為複雜、不容易把握的文字訊息，試試看能否轉換成「分數」的概念。因此，本次報告的題目是想找出分數和文字評論之間的關聯，希望能在看到文字訊息時，若要轉換成分數，能有一些參考項目。

二、假設及猜想

如果說分數高低代表的是人們對於餐廳的情緒的話，就一般的情況來看，人們情緒達到極端的時候，不論好壞，好像都會比較想把情緒發洩出來，而這反映在對事物的評論中，感覺會比較用心去評論，用不用心最直接的反應就是在句子的長短中，再來可能是句子中語彙的豐富程度，如果假設成立的話，那麼分數越高或越低的時候，句子以及語彙應該也越豐富等。當然，生氣的時候，也可能只是罵一句髒話就結束，所以分數低的情況，因此，句子也可能是最短的也不一定。

三、資料來源以及資料前處理

資料來源於 kaggle，先刪去用不到的欄位：['Pictures']、['7514']、['Metadata']，然後在創立需要用到的三個欄位，分別是計算句子長度的[Reviews_len]，以及語彙豐富度(不同詞彙/長度)的[Reviews_Diversity]、還有評論情緒的[Reviews_emotion]。

處理後的資料表格：

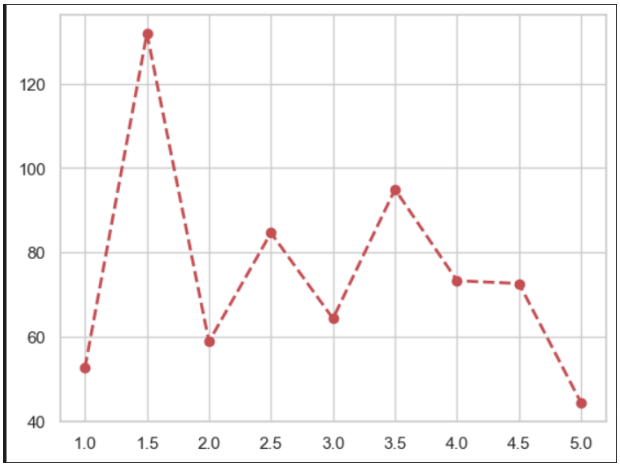
	Restaurant	Reviewer	Review	Rating	Time	Reviews_len	Reviews_Diversity	Reviews_emotion
0	Beyond Flavours	Rusha Chakraborty	The ambience was good, food was quite good . h...	5.0	5/25/2019 15:54	45	0.144144	0.9664
1	Beyond Flavours	Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. S...	5.0	5/25/2019 14:20	31	0.208333	0.9360
2	Beyond Flavours	Ashok Shekhawat	A must try.. great food great ambience. Thnx f...	5.0	5/24/2019 22:54	37	0.169312	0.9186
3	Beyond Flavours	Swapnil Sarkar	Soumen das and Arun was a great guy. Only beca...	5.0	5/24/2019 22:11	32	0.195946	0.8591
4	Beyond Flavours	Dileep	Food is good.we ordered Kodi drumsticks and ba...	5.0	5/24/2019 21:37	33	0.193750	0.9201

（餐廳和評論的關係，大多都在 3.5-4 分之間）

（文字雲）

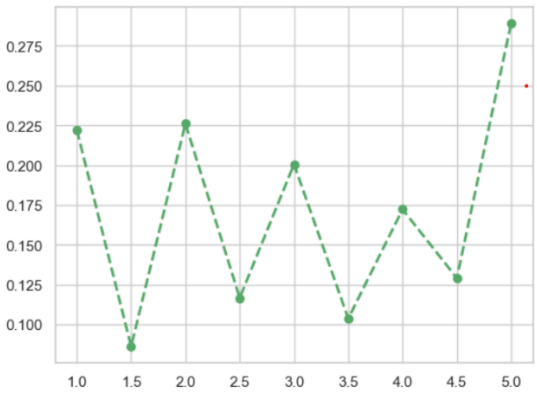
四、結果呈現及圖表分析

1. 評論長度和分數之間的關係



透過折線圖我們可以看到，在評論和長度的關係之間，大部分的評論都不會超過 100 字，除了「1.5」分的這一欄，但這個項目的資料數太少，因此不太適合作為採用項目，而除了這一欄以外，其他欄位的字數長度也都不超過一百，而且分數高低和字數長度也並沒有明確的關聯，雖然字數長度有高有低，但彼此的誤差也僅在 10 個字左右而已。所以句子的長度，其實不太有助於反映人們心中的好惡。

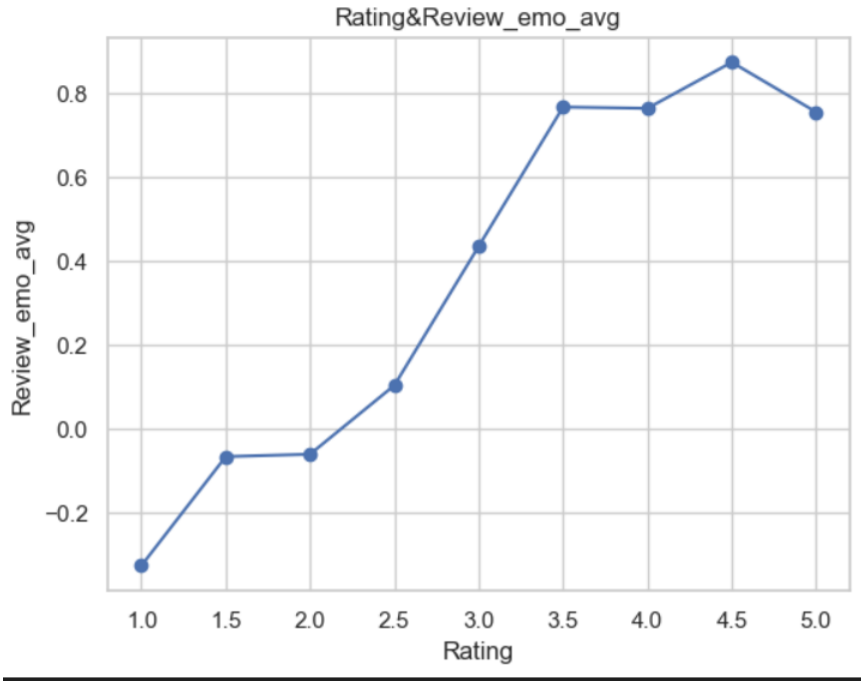
2. 語彙豐富度和評論的關係



透過折線圖我們能夠看到，除了「1.5」分這欄因為資料數目比較少，所可能導致的結果有偏差以外，其他分數的數值差距並不大，除了五分以外，其他欄位的分數高低也和語彙豐富度沒有呈現正相關，只有在五分的情況下，語彙豐富度才有比較明顯的區別。

因此，語意豐富度在 5.0 分以下的區別作用並不強，所以可以做一個推論是如果一篇評論能讓人感覺到它所使用語彙非常的豐富的話，那麼他的評論分數應該會是非常高才是。

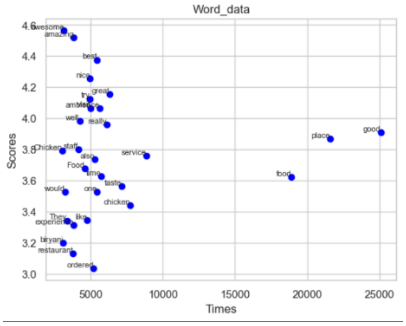
3. 情緒和評論的關係



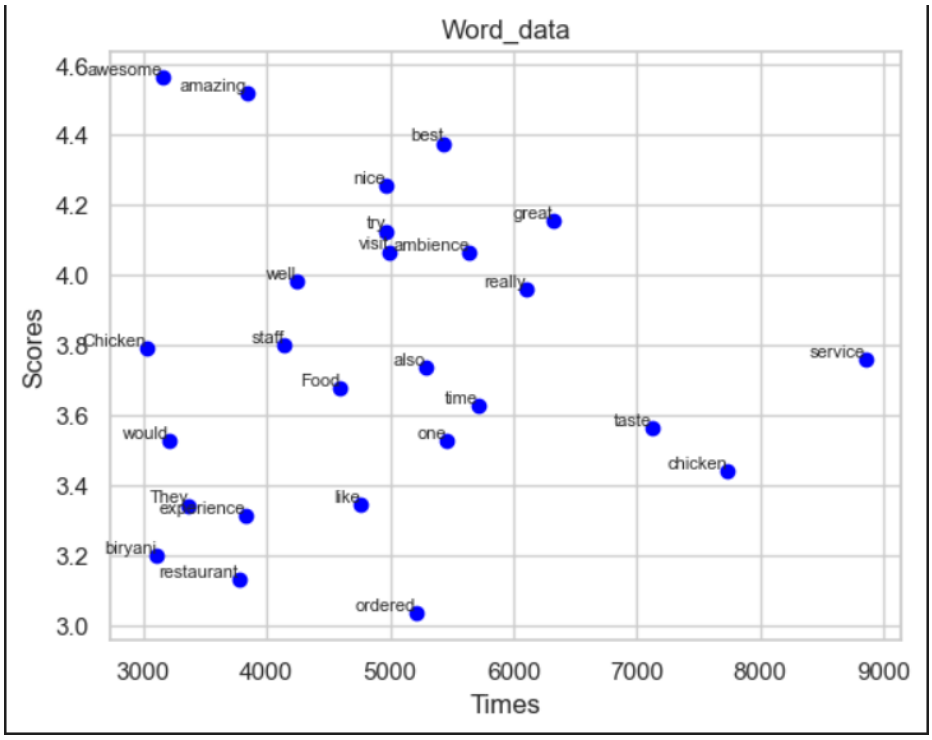
透過圖表我們可以知曉，情緒和分數的關聯確實是呈現正相關的，不過在 2.5-3.5 的這個區間，線的斜率較高，而在這個區間以外的，不論高低，就呈現比較平緩的樣態。

所以評論的確和評論中所呈現的情緒有正相關，但是在區好及更好、壞及更壞的作用中並不大，不過情緒確實能夠很好的區別「好、壞」。

4. 字詞和分數的關係



（因為另外三個有點影響到圖，所以又額外在繪製一張）



透過這張圖我們可以知道，那些本來就帶有稱讚意味的詞，的確代表著比較高的分數，但是那些單純表意義較無好壞判斷的名詞，如：Ambience、service、taste、food、place 等，也代表著偏高的分數，我推測可能的原因是因為這些詞雖然在評論好壞的時候都會用到，但會特別提到這些項目，通常就已經代表著評論者對餐廳有比較好的印象，所以才會伴隨著比較高的分數。

五、結論

句子長度和評論的關係沒有想像中的大，前面假設本身的猜想就不是很準確以外，再者可能是因為長度所牽涉的問題太多了，例如中文可能會說的「天壽讚」之類的評論，雖然短短的，但也象徵極高的評價，而這個也可以用來對照 5.0 分數的句子是最短的現象。詞彙豐富度在分數極高的情況中會被凸顯出來，因此可以當作一個用來判斷好的依據，而可能的原因我覺得也許和字數長度有關，因為長度比較短，所以語彙豐富度自然會有比較高的數值。而情緒分析的確是一個很好用來區別好壞的指標，但是它的侷限性在於說無法對於好壞做更進一步的區別，較只能用來做一個比較廣泛的劃分，此外，額外發現有些僅僅只是表現餐廳的內容的名詞，也能作為參考的依據。